PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12Q 1/68, G06F 17/30, 159/00

A1

(11) International Publication Number: WO 99/05323

(43) International Publication Date: 4 February 1999 (04.02.99)

(21) International Application Number: PCT/US98/15151

(22) International Filing Date: 24 July 1998 (24.07.98)

(30) Priority Data:

60/053,842	25 July 1997 (25.07.97)	US
60/069,198	11 December 1997 (11.12.97)	US
60/069,436	11 December 1997 (11.12.97)	US

(71) Applicant: AFFYMETRIX, INC. [US/US]; 3380 Central Expressway, Santa Clara, CA 95051 (US).

(72) Inventor: BALABAN, David, J.; 37 Bret Harte Road, San Rafael, CA 94901 (US).

(74) Agents: LANG, Dan, H. et al.; Townsend and Townsend and Crew LLP, 8th floor, Two Embarcadero Center, San Francisco, CA 94111-3834 (US).

(81) Designated States: JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: GENE EXPRESSION AND EVALUATION SYSTEM

(57) Abstract

An efficient and easy to use query system for a gene expression database. Using such a system, one can easily identify genes or expressed sequence tags whose expression correlates to particular tissue types. Various tissue types may correspond to different diseases, states of disease progression, different organs, different species, etc. Researchers may now use large scale gene expression databases to full advantage.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

GENE EXPRESSION AND EVALUATION SYSTEM

CROSS-REFERENCE TO RELATED APPLICATIONS

The present application claims priority from U.S. Prov. App. No. 60/053,842 filed July 25, 1997, entitled COMPREHENSIVE BIO-INFORMATICS DATABASE, from U.S. Prov. App. No. 60/069,198 filed on December 11, 1997, entitled COMPREHENSIVE DATABASE FOR BIOINFORMATICS , and from U.S. Prov. App. No. 60/069,436, entitled GENE EXPRESSION AND EVALUATION SYSTEM, filed on December 11, 1997. The contents of all three provisional applications are herein incorporated by reference.

The subject matter of the present application is related to the subject matter of the following three co-assigned applications filed on the same day as the present application: METHOD AND APPARATUS FOR PROVIDING A BIOINFORMATICS DATABASE (Attorney Docket No. 018547-033810), METHOD AND SYSTEM FOR PROVIDING A POLYMORPHISM DATABASE (Attorney Docket No. 018547-033820), METHOD AND SYSTEM FOR PROVIDING A PROBE ARRAY CHIP DESIGN DATABASE (Attorney Docket No. 018547-033830). The contents of these three applications are herein incorporated by reference.

BACKGROUND OF THE INVENTION

The present invention relates to computer systems and more particularly to computer systems for analyzing expression levels or concentrations.

Devices and computer systems have been developed for collecting information about gene expression or expressed sequence tag (EST) expression in large numbers of tissue samples. For example, PCT application WO92/10588, incorporated herein by reference for all purposes, describes techniques for sequencing or sequence checking nucleic acids and other materials. Probes for performing these operations may be formed in arrays according to the methods of, for example, the pioneering techniques disclosed in U.S. Patent No. 5,143,854 and U.S. Patent No. 5,571,639, both incorporated herein by reference for all purposes.

According to one aspect of the techniques described therein, an array of nucleic acid probes is fabricated at known locations on a chip or substrate. A fluorescently

labeled nucleic acid is then brought into contact with the chip and a scanner generates an image file indicating the locations where the labeled nucleic acids bound to the chip. Based upon the identities of the probes at these locations, it becomes possible to extract information such as the monomer sequence of DNA or RNA.

Computer-aided techniques for monitoring gene expression using such arrays of probes have been developed as disclosed in EP Pub. No. 0848067 and PCT publication No. WO 97/10365, the contents of which are herein incorporated by reference. Many disease states are characterized by differences in the expression levels of various genes either through changes in the copy number of the genetic DNA or through changes in levels of transcription (*e.g.*, through control of initiation, provision of RNA precursors, RNA processing, etc.) of particular genes. For example, losses and gains of genetic material play an important role in malignant transformation and progression. Furthermore, changes in the expression (transcription) levels of particular genes (*e.g.*, oncogenes or tumor suppressors), serve as signposts for the presence and progression of various cancers.

Information on expression of genes or expressed sequence tags may be collected on a large scale in many ways, including the probe array techniques described above. One of the objectives in collecting this information is the identification of genes or ESTs whose expression is of particular importance. Researchers wish to answer questions such as: 1) Which genes are expressed in cells of a malignant tumor but not expressed in either healthy tissue or tissue treated according to a particular regime? 2) Which genes or ESTs are expressed in particular organs but not in others? 3) Which genes or ESTs are expressed in particular species but not in others?.

Collecting vast amounts of expression data from large numbers of samples including all the tissue types mentioned above is but the first step in answering these questions. To derive full value from the investment made in collecting and storing expression data, one must be able to efficiently mine the data to find items of particular relevance. What is needed is an efficient and easy to use query system for a gene expression database.

SUMMARY OF THE INVENTION

An efficient and easy to use query system for a gene expression database is provided by virtue of the present invention. Using such a system, one can easily identify

genes or expressed sequence tags whose expression correlates to particular tissue types. Various tissue types may correspond to different diseases, states of disease progression, different organs, different species, etc. Researchers may now use large scale gene expression databases to full advantage.

According to a first aspect of the present invention, a method is provided in a computer system for operating a database storing information about compound concentration. The method includes: providing a database including concentrations of a plurality of compounds as measured in a plurality of samples, accepting a user query to the database to identify desired ones of the plurality of compounds, the user query specifying concentration characteristics of the desired compounds in selected ones of the plurality of samples, and comparing the concentration characteristics to the concentrations stored in the database to identify the desired compounds.

A further understanding of the nature and advantages of the inventions herein may be realized by reference to the remaining portions of the specification and the attached drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 illustrates an example of a computer system that may be used to execute software embodiments of the present invention.

Fig. 2 shows a system block diagram of a typical computer system.

Fig. 3 is a flowchart describing steps of developing expression data according to one embodiment of the present invention.

Fig. 4 is a flowchart describing steps of querying an expression database according to one embodiment of the present invention.

Figs. 5A-5L depict a user interface for querying an expression database according to one embodiment of the present invention.

DESCRIPTION OF SPECIFIC EMBODIMENTS

Fig. 1 illustrates an example of a computer system that may be used to execute software embodiments of the present invention. Fig. 1 shows a computer system 1 which includes a monitor 3, screen 5, cabinet 7, keyboard 9, and mouse 11. Mouse 11 may have

one or more buttons such as mouse buttons 13. Cabinet 7 houses a CD-ROM drive 15 and a hard drive (not shown) that may be utilized to store and retrieve software programs including computer code incorporating the present invention. Although a CD-ROM 17 is shown as the computer readable medium, other computer readable media including floppy disks, DRAM, hard drives, flash memory, tape, and the like may be utilized. Cabinet 7 also houses familiar computer components (not shown) such as a processor, memory, and the like.

Fig. 2 shows a system block diagram of computer system 1 used to execute software embodiments of the present invention. As in Fig. 1, computer system 1 includes monitor 3 and keyboard 9. Computer system 1 further includes subsystems such as a central processor 50, system memory 52, I/O controller 54, display adapter 56, removable disk 58, fixed disk 60, network interface 62, and speaker 64. Removable disk 58 is representative of removable computer readable media like floppies, tape, CD-ROM, removable hard drive, flash memory, and the like. Fixed disk 60 is representative of an internal hard drive or the like. Other computer systems suitable for use with the present invention may include additional or fewer subsystems. For example, another computer system could include more than one processor 50 (*i.e.*, a multi-processor system) or memory cache.

Arrows such as 66 represent the system bus architecture of computer system 1. However, these arrows are illustrative of any interconnection scheme serving to link the subsystems. For example, display adapter 56 may be connected to central processor 50 through a local bus or the system may include a memory cache. Computer system 1 shown in Fig. 2 is but an example of a computer system suitable for use with the present invention. Other configurations of subsystems suitable for use with the present invention will be readily apparent to one of ordinary skill in the art. In one embodiment, the computer system is an IBM compatible personal computer.

The VLSIPS™ and GeneChip™ technologies provide methods of making and using very large arrays of polymers, such as nucleic acids, on very small chips. *See* U.S. Patent No. 5,143,854 and PCT Patent Publication Nos. WO 90/15070 and 92/10092, each of which is hereby incorporated by reference for all purposes. Nucleic acid probes on the chip are used to detect complementary nucleic acid sequences in a sample nucleic acid of interest (the "target" nucleic acid).

It should be understood that the probes need not be nucleic acid probes but

may also be other polymers such as peptides. Peptide probes may be used to detect the concentration of peptides, polypeptides, or polymers in a sample. The probes should be carefully selected to have bonding affinity to the compound whose concentration they are to be used to measure.

In one embodiment, the present invention provides methods of reviewing and analyzing information relating to the concentration of compounds in a sample as measured by monitoring affinity of the compounds to polymers such as polymer probes. In a particular application, the concentration information is generated by analysis of hybridization intensity files for a chip containing hybridized nucleic acid probes. The hybridization of a nucleic acid sample to certain probes may represent the expression level of one more genes or expressed sequence tags (EST). The expression level of a gene or EST is herein understood to be the concentration within a sample of mRNA or protein that would result from the transcription of the gene or EST.

Expression level information that is reviewed and/or analyzed by virtue of the present invention need not be obtained from probes but may originate from any source. If the expression information is collected from a probe array, the probe array need not meet any particular criteria for size and density. Furthermore, the present invention is not limited to reviewing and/or analyzing fluorescent measurements of bondings such as hybridizations but may be readily utilized for reviewing and/or analyzing other measurements.

Concentration of compounds other than nucleic acids may be reviewed and/or analyzed according to one embodiment of the present invention. For example, a probe array may include peptide probes which may be exposed to protein samples, polypeptide samples, or peptide samples which may or may not bond to the peptide probes. By appropriate selection of the peptide probes, one may detect the presence or absence of particular proteins, polypeptides, or peptides which would bond to the peptide probes.

A system that designs a chip mask, synthesizes the probes on the chip, labels nucleic acids from a target sample, and scans the hybridized probes is set forth in U.S. Patent No. 5,571,639 which is hereby incorporated by reference for all purposes. However, the present invention may be used separately for reviewing and/or analyzing the results of other systems for generating expression information, or for reviewing and/or analyzing concentrations of polymers other than nucleic acids.

The term "perfect match probe" refers to a probe that has a sequence that is perfectly complementary to a particular target sequence. The test probe is typically perfectly complementary to a portion (subsequence) of the target sequence. The term "mismatch control" or "mismatch probe" refer to probes whose sequence is deliberately selected not to be perfectly complementary to a particular target sequence. For each mismatch (MM) control in an array there typically exists a corresponding perfect match (PM) probe that is perfectly complementary to the same particular target sequence.

Among the important pieces of information obtained from the chips are the relative fluorescent intensities obtained from the perfect match probes and mismatch probes. These intensity levels are used to estimate an expression level for a gene or EST. The computer system used for analysis will preferably have available other details of the experiment including possibly the gene name, gene sequence, probe sequences, probe locations on the substrate, and the like.

An expression analysis is performed for each gene for each experiment. Fig. 3 is a flowchart describing steps of estimating an expression level for a particular gene as measured in a particular experiment on a chip. At step 302, the computer system receives raw scan data of N pairs of perfect match and mismatch probes. In a preferred embodiment, the hybridization intensities are photon counts from a fluorescein labeled target that has hybridized to the probes on the substrate. For simplicity, the hybridization intensity of a perfect match probe will be designed "I_{pm}" and the hybridization intensity of a mismatch probe will be designed "I_{mm}."

Hybridization intensities for a pair of probes are retrieved at step 304. The background signal intensity is subtracted from each of the hybridization intensities of the pair at step 306. Background subtraction can also be performed on all the raw scan data at the same time.

At step 308, the hybridization intensities of the pair of probes are compared to a difference threshold (D) and a ratio threshold (R). It is determined if the difference between the hybridization intensities of the pair ($I_{pm} - I_{mm}$) is greater than or equal to the difference threshold AND the quotient of the hybridization intensities of the pair (I_{pm} / I_{mm}) is greater than or equal to the ratio threshold. The difference thresholds are typically user defined values that have been determined to produce accurate expression monitoring of a gene or

genes. In one embodiment, the difference threshold is 20 and the ratio threshold is 1.2.

If $I_{pm} - I_{mm} >= D$ and $I_{pm} / I_{mm} >= R$, the value NPOS is incremented at step 310. In general, NPOS is a value that indicates the number of pairs of probes which have hybridization intensities indicating that the gene is likely expressed. NPOS is utilized in a determination of the expression of the gene.

At step 312, it is determined if $I_{mm} - I_{pm} >= D$ and $I_{mm} / I_{pm} >= R$. If these expressions are true, the value NNEG is incremented at step 314. In general, NNEG is a value that indicates the number of pairs of probes which have hybridization intensities indicating that the gene is likely not expressed. NNEG, like NPOS, is utilized in a determination of the expression of the gene.

For each pair that exhibits hybridization intensities either indicating the gene is expressed or not expressed, a log ratio value (LR) and intensity difference value (IDIF) are calculated at step 316. LR is calculated by the log of the quotient of the hybridization intensities of the pair (I_{pm} / I_{mm}). The IDIF is calculated by the difference between the hybridization intensities of the pair ($I_{pm} - I_{mm}$). If there is a next pair of hybridization intensities at step 318, they are retrieved at step 304.

For each analysis performed certain data is stored in an expression analysis database. There is preferably a record for each gene or EST for which the chip measures expression. This record includes fields to hold various pieces of information. One field stores an analysis ID to identify the analysis. A result type ID field indicates whether the listed expression results indicate that the gene is present, marginal, absent, or unknown based on application of a decision matrix to the values P1, P2, P3, and P4. A number_positive field shows NPOS. An number_negative field shows NNEG. A number_used field shows the number of probes belonging to pairs that incremented NNEG or NPOS. A number_all field indicates N. An average log ratio field indicates the average LR for all probe pairs. A number_positive_exceeds field indicates the value of NPOS - NNEG. A number_negative_exceeds field indicates the value of NNEG - NPOS. An average differential intensity field indicates the average IDIF for the probe pairs. A number_in_average field indicates the number of probe pairs used in computing the average.

Steps of operating a user interface to the expression database will now be illustrated with reference to Fig. 4. The steps of Fig. 4 may be repeated or may occur in a

different order, or one or more steps may be omitted. The discussion of the user interface will also refer to Figs. 5A-5L which depict representative screen displays of the user interface.

At step 402, the user selects files of expression analysis results for querying. Fig. 5A illustrates an interface screen where the user may specify expression results files. Each file represents one experiment. A table 502 lists the files that have already been selected. A given list may be saved for later use by selecting a button 504. A previously saved list may be deleted by selecting a button 506. A button 508 resets the list depicted in table 502 to a previously saved version. An import button 512 imports the contents of the files depicted in table 502 for querying. Within table 502, a file name column lists the file names that would be imported by application of import button 512. A code column indicates the tissue type for the expression data in each file. A replicate file indicates whether the file is a duplicate. A chip design code column indicates the chip design used to generate the data for the file. Various other columns (not shown) give further information about the analysis result data.

By selecting a select files button 514, the user calls up a select files screen 516 as shown in Fig. 5B. This provides an interactive file search and selection process that does not require typing in the file name. Before importing the file list, the user should select a species by using a species drop-down list 518 as shown in Fig. 5C. An analysis-type drop down list 519 allows the user to select between a relative expression analysis and an absolute expression analysis.

Fig. 5D shows a normalization form 520 for normalizing imported expression results at step 404. The software scales the average difference data generated by the analysis routine based on the user's selections on normalization form 520. In a chip variability area 522, the user specifies housekeeping genes with known expression levels and selects a scale value. The user can elect to either apply or not apply this scale value. If the user elects to apply the scale value, each gene expression level measured on a single chip is multiplied by a value equal to the desired scaling factor divided by the average of housekeeping expression levels measured on that chip.

Also on normalization form 520, in a tissue variability area 524, the user may select a scale value that applies to data collected from multiple chips and whether or not it is

applied. If this scale value is to be applied, each expression value measured in a chip set is multiplied by a factor equal to the scale value divided by the average expression level measured over all genes for the entire chip set. A transformation area 526 allows the user to select whether negative average difference values are to be converted to positive numbers by use of a logarithmic transform. The user can reset all the changes made on normalization form 520 by selecting a reset button 528 or apply the selected normalizations and transformations by selecting an apply button 530.

At step 406, the user filters the large set of experimental data that was imported, normalized, and transformed. Fig. 5E depicts a filter experiments form 532. A lower table 534 lists the imported experiments and genes or EST and the expression data associated with each combination of experiment and gene or EST. An upper table 536 is used to enter a query to filter the experiment data in lower table 534. Each column of upper table 536 corresponds to a column in lower column 534. Upper table 536 is similar to a query by example (QBE) grid as included in Microsoft Access. Predicates are entered in the columns of upper table 536 with all the predicates in a single row treated as ANDs and those between rows treated as OR's. The results satisfying a given query are displayed in lower table 534 upon selection of a filter button 538. Filters may be saved, deleted, and reset by use of appropriately labeled buttons, 540, 542, and 544. A stored filter may be loaded by use of a drop-down list 546. Selection of an export button 547 writes the data to an Exel spreadsheet.

To facilitate further user queries, the user may specify a new field to be used as a pivot field for future queries at step 408. Elements of the selected field will become columns in the new table. Fig. 5F shows how a pivot value is selected by use of a drop-down list 548. The pivot value identifies the expression data that will be listed in the columns of lower table 534. Fig. 5G shows a pivot column drop-down list 550 allows selection of a particular column of lower table 534 as the pivot field. The entries of the selected column are shown in a left list box 552 and moved to a right list box 554 to include them as rows in the pivoted table. The user selects arrow keys 556 to add and delete items of right list box 554. To perform the pivot operation, the user selects a pivot button 558.

Fig. 5H depicts a user interface for filtering tissue types as displayed as a result of the pivot operation. Lower table 534 shows the result of a pivot operation as

described with reference to Figs. 5F-5G.

Upper table 536 is now used at step 410 to specify a query to filter genes using the results of experiments obtained from different tissue types. Again, predicates in a row are treated as ANDs. Predicates between rows are treated as ORs. By properly formulating a query, the user may answer questions such as which genes are up-regulated in normal tissue and down-regulated in diseased tissue. The depicted Entrez definition column contains the definition column from the public domain Entrez database. The depicted query marked 'like "growth"' retains those records having the string "growth" as a substring in the designated column.

One condition satisfying the depicted query is that a gene have an expression level in experiment 4002736D greater than 10 and an expression level in experiment 4003228A greater than 10 and less than 0.6 times the expression level in experiment 4002736D. An alternate condition satisfying the query is that the expression level in experiment 4002736D be greater than 10 and the expression level in experiment 4003228A greater than 10 and greater than 1.4 times the expression level in experiment 4002736D.

This query determines the genes that have a particular fold change pattern between experiment 4003228A and experiment 4002736D. It will filter out genes for which there is no significant fold change between the experiments. Specifically, it finds all genes for which the expression level of experiment 4003228A is less than 60% of the expression level of experiment 4002736D, or for which the expression level of experiment 4003228A is greater than 140% of the expression level of experiment 4002736D. Both experiments are also constrained to have expression levels greater than 10.

Filters may be saved or reset by selection of buttons 560 and 562, respectively. The records displayed in lower table 534 may be sorted on any column(s), and columns may be hidden, frozen, or repositioned for better viewing. Lower table 534 may also be saved in different formats, including a spreadsheet format such as Microsoft Excel, by clicking on an export button 564. A saved filter may be accessed via a pull down menu 566 or deleted by selection of a delete button 568. Additional information on any gene may be obtained by double clicking its row. This will load an Internet browser program and open a web site such as the Entrez web site that stores information for the gene. The browser program then displays the entry for that gene.

At step 412, by selecting a graph button 570, the user calls up a scatter-plot display 572 depicted in Fig. 5I. Two experiments are selected for comparison using drop-down lists 574 and 576 for the x axis and y axis respectively. The graph is generated by selecting a build scatter button 578. Each point on the scatter plot corresponds to a particular gene. The point is positioned on the graph according to its measured expression level in both experiments. By checking a box 580, the user may select to have the points color coded according to whether the gene was present in both (2P), one (1P), or neither (OP) of the experiments. By checking one or more of boxes 582, the user may elect to show or not show genes according to this categorization.

By making an appropriate selection in a box 584, the user may select an interpretation for future mouse clicks. One choice is for the system to do nothing in response to a mouse click. Another choice is for the system to show gene data for a point selected by a mouse click. The gene data appears in a box 586 including the accession number, the gene name, the expression levels as measured in a variety of experiments, and an expression call for each experiment (either absent or present.) An Entrez definition name is also shown. Double clicking on an entry will invoke an Internet browser to show the Entrez entry for the gene.

The user may also select "rope" in box 584 to collect interesting points for comparison by surrounding them with a polygon. Lines are automatically drawn between each mouse click, encircling those genes to be included in a bar graph. The user may display the bar graph by selecting a button 588.

At step 414, Fig. 5J depicts a bar graph 590 for the roped genes in the scatter plot of Fig. 5I. Each grouping of bars in Fig. 5J corresponds to a gene. Each bar within a grouping corresponds to an experiment and is color-coded according to a legend 592. Initially only two experiments are displayed, the two experiments corresponding to the axes of the scatter plot of Fig. 5I. However, the user may select further experiments from a box 594. Once the desired experiments are selected, the user selects a build button 596 to display the desired bar graph. A table 598 shows the expression levels for the depicted genes.

For the display of Fig. 5J, the option "gene" is selected in a box 600. To view individual plots of the expression level for each gene as they vary over the experiments, the user may select option "experiment" in box 600 before selecting build button 596. This

produces a line graph 602 as shown in Fig. 5K. The experiments are arranged along the horizontal axis in the order specified in box 594. Each gene has its own trace corresponding to its expression level as it varies over the experiments. A legend 604 identifies the trace for each gene. To change the position of an experiment along the horizontal axis, the user uses up and down arrows 606 and 608 to change its position. This feature makes it possible to reorder the experiments to reflect additional sequencing knowledge. For example, if the experiments represent a time course such as progression of a disease or treatment, they can be graphically ordered in time sequence. The graph then represents the change in expression level as a function of time for the selected gene. A slider icon 612 allows the user to scroll along the horizontal axis if line graph 602 does not fit on the screen. A maker check box 614 shows a horizontal line across line graph 602 defining a particular expression level. This allows the user to easily view data points above the selected level.

More information about a gene may be obtained by clicking on any bar in the group. All of the information for the gene will be displayed in a separate window 610 as shown in Fig. 5L.

In the foregoing specification, the invention has been described with reference to specific exemplary embodiments thereof. It will, however, be evident that various modifications and changes may be made thereunto without departing from the broader spirit and scope of the invention as set forth in the appended claims and their full scope of equivalents. For example, it will be understood that wherever "expression level" is referred to, one may substitute the measured concentration of any compound. Also, wherever "gene" is referred to, one may substitute the term "expressed sequence tag."

WHAT IS CLAIMED IS:

1. In a computer system, a method for operating a database storing expression level information comprising:

providing a database comprising expression levels for each of a plurality of genes or expressed sequence tags (EST) as measured in each of a plurality of tissue types;

accepting a user query to said database to identify desired ones of said plurality of genes or EST, said user query specifying expression level characteristics of said desired genes; and

comparing said expression level characteristics to said expression levels stored in said database to identify said desired genes or EST.

2. The method of claim 1 further comprising:

displaying information identifying said desired genes or EST.

3. The method of claim 1 wherein said plurality of tissue types comprise a diseased tissue type.

4. The method of claim 1 wherein said plurality of tissue types comprise a healthy tissue type.

5. The method of claim 1 wherein said plurality of tissue types comprise a cancerous tissue type.

6. The method of claim 1 wherein said plurality of tissue types comprise a drug treated tissue type.

7. The method of claim 1 wherein said plurality of tissue types comprise issues obtained from disparate species.

8. The method of claim 1 wherein said plurality of tissue types comprise tissues obtained from disparate organs.

9. The method of claim 1 wherein said expression level characteristics comprise expression level ranges as measured for a particular gene in at least two of said plurality of tissue types.

10. The method of claim 1 wherein said expression level characteristics comprise relationships among expression levels as measured for a particular gene in at least two of said plurality of tissue types.

11. The method of claim 1 further comprising:

accepting user input selecting two of said plurality tissue types for graphical display;

displaying a first axis corresponding to a first one of said two tissue types;

displaying a second axis corresponding to a second one of said two tissue types;

for a selected one of said plurality of genes or EST, displaying a mark at a position wherein said position is selected relative to said first axis in accordance with an expression level of said selected gene or EST measured in said first tissue type and selected relative to said second axis in accordance with an expression level of said selected gene or EST measured in said second tissue type.

12. The method of claim 11 further comprising:

repeating said operation of displaying a mark for a plurality of selected genes or EST.

13. In a computer system, a method for operating a database storing information about compound concentration comprising:

providing a database comprising concentrations of a plurality of compounds as measured in a plurality of samples;

accepting a user query to said database to identify desired ones of said plurality of compounds, said user query specifying concentration characteristics of said

desired compounds in selected ones of said plurality of samples; and

comparing said concentration characteristics to said concentrations stored in said database to identify said desired compounds.

14. A computer program product for operating a database storing expression level information comprising:

code that provides a database comprising expression levels for each of a plurality of genes or expressed sequence tags (EST) as measured in each of a plurality of tissue types;

code that accepts a user query to said database to identify desired ones of said plurality of genes or EST, said user query specifying expression level characteristics of said desired genes;

code that compares said expression level characteristics to said expression levels stored in said database to identify said desired genes or EST; and

a computer-readable storage medium for storing the codes.

15. The product of claim 14 further comprising:

code that displays information identifying said desired genes or EST.

16. The product of claim 14 wherein said plurality of tissue types comprise a diseased tissue type.

17. The product of claim 14 wherein said plurality of tissue types comprise a healthy tissue type.

18. The product of claim 14 wherein said plurality of tissue types comprise a cancerous tissue type.

19. The product of claim 14 wherein said plurality of tissue types comprise a drug treated tissue type.

20. The product of claim 14 wherein said plurality of tissue types comprise tissues obtained from disparate species.

21. The product of claim 14 wherein said plurality of tissue types comprise tissues obtained from disparate organs.

22. The product of claim 14 wherein said expression level characteristics comprise expression level ranges as measured for a particular gene in at least two of said plurality of tissue types.

23. The product of claim 14 wherein said expression level characteristics comprise relationships among expression levels as measured for a particular gene in at least two of said plurality of tissue types.

24. The product of claim 14 further comprising:

code that accepts user input selecting two of said plurality tissue types for graphical display;

code that displays a first axis corresponding to a first one of said two tissue types;

code that displays a second axis corresponding to a second one of said two tissue types;

code that, for a selected one of said plurality of genes or EST, displays a mark at a position wherein said position is selected relative to said first axis in accordance with an expression level of said selected gene or EST measured in said first tissue type and selected relative to said second axis in accordance with an expression level of said selected gene or EST measured in said second tissue type.

25. The product of claim 24 further comprising:

code that repeatedly applies said code that displays a mark for a plurality of selected genes or EST.

26. A computer program product for operating a database storing information about compound concentration comprising:

code that receives a database comprising concentrations of a plurality of compounds as measured in a plurality of samples;

code that accepts a user query to said database to identify desired ones of said plurality of compounds, said user query specifying concentration characteristics of said desired compounds in selected ones of said plurality of samples; and

code that compares said concentration characteristics to said concentrations stored in said database to identify said desired compounds.

27. A computer system comprising:

a processor; and

a memory storing code to operate said processor, said code comprising:

code that provides a database comprising expression levels for each of a plurality of genes or expressed sequence tags (EST) as measured in each of a plurality of tissue types;

code that accepts a user query to said database to identify desired ones of said plurality of genes or EST, said user query specifying expression level characteristics of said desired genes; and

code that compares said expression level characteristics to said expression levels stored in said database to identify said desired genes or EST.

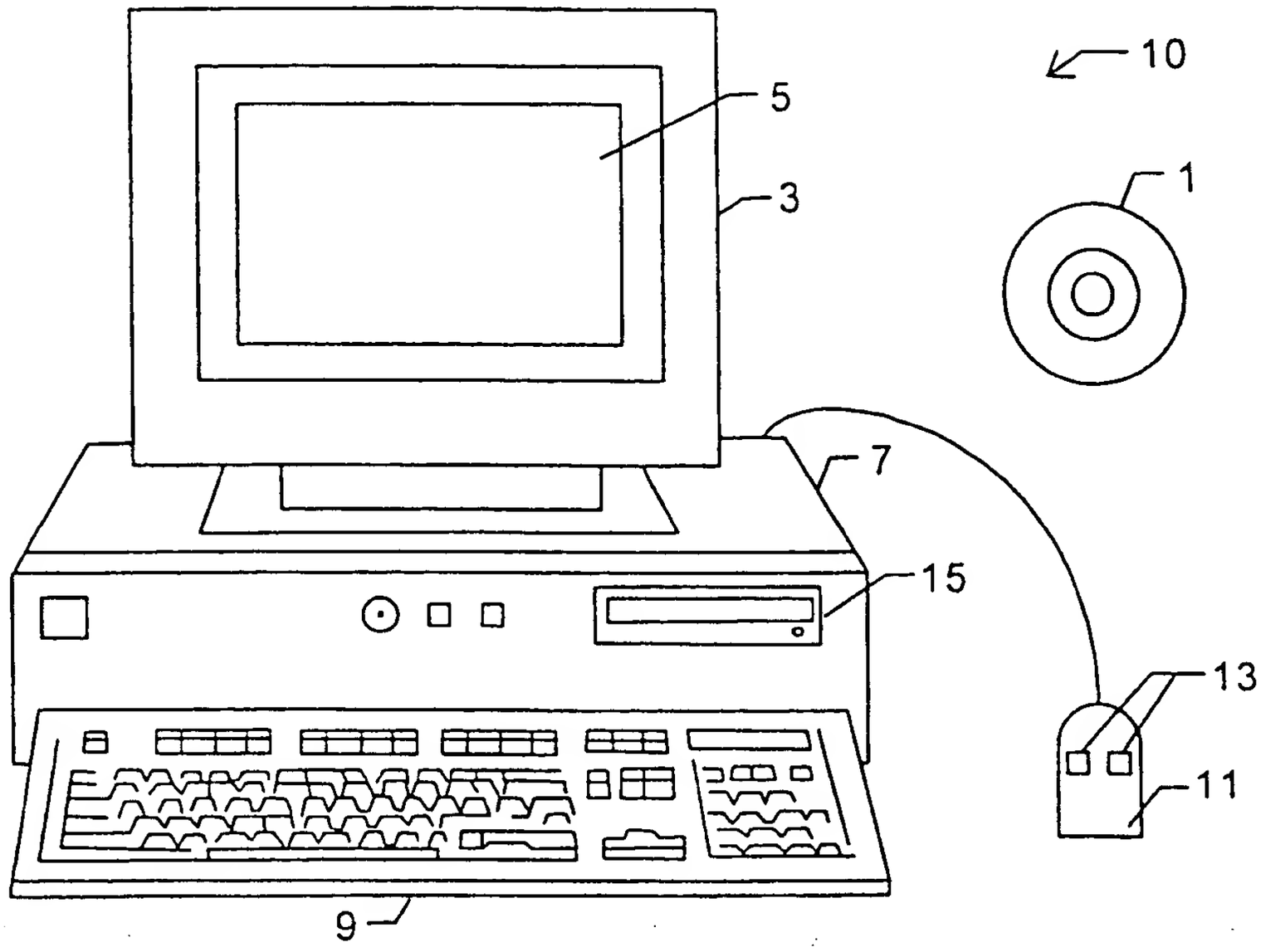
10
5
1
3
7
15
13
11
9

FIG. 1

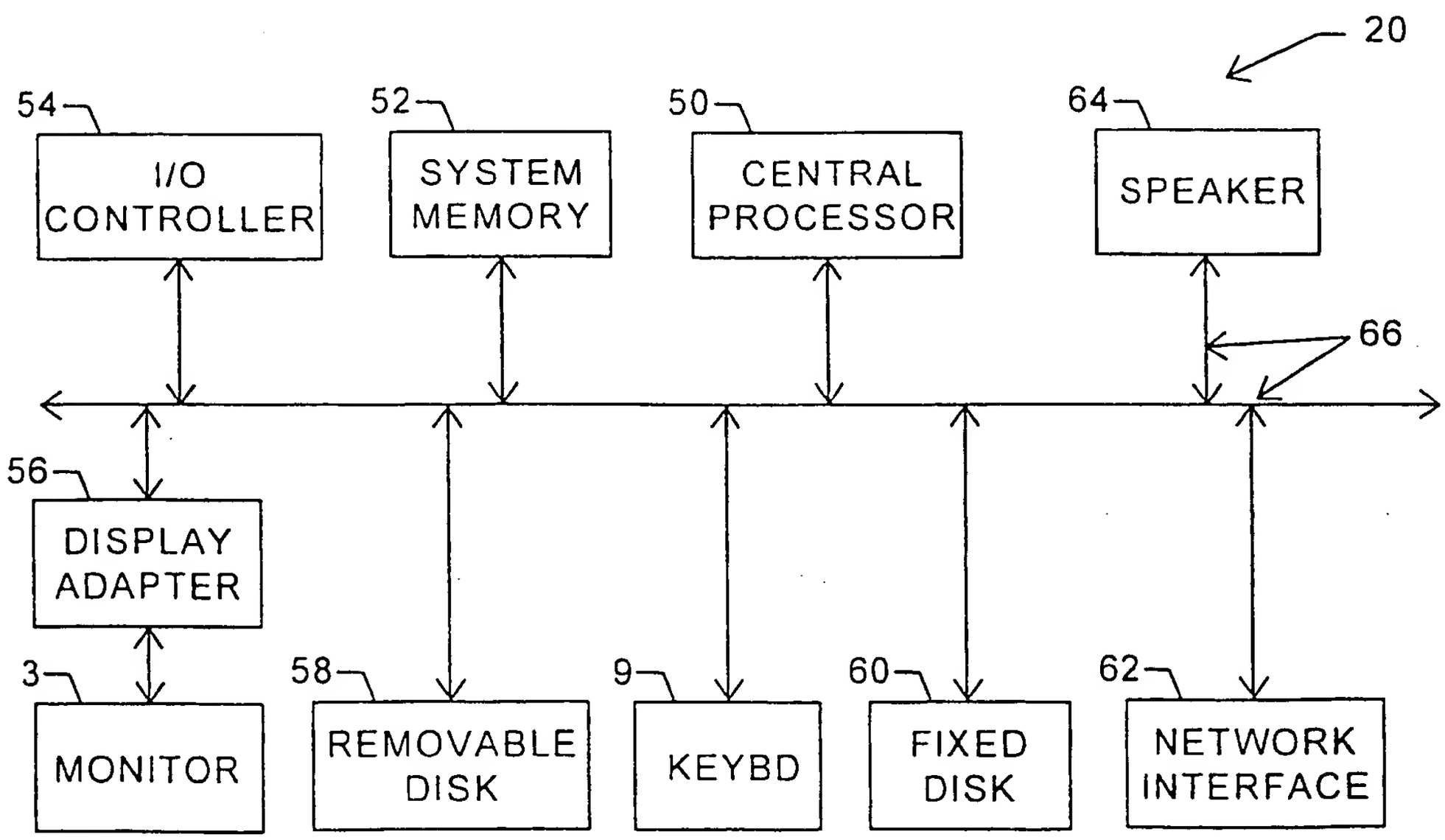
20
54
52
50
64
I/O CONTROLLER
SYSTEM MEMORY
CENTRAL PROCESSOR
SPEAKER
66
56
DISPLAY ADAPTER
3
MONITOR
58
REMOVABLE DISK
9
KEYBD
60
FIXED DISK
62
NETWORK INTERFACE

FIG. 2

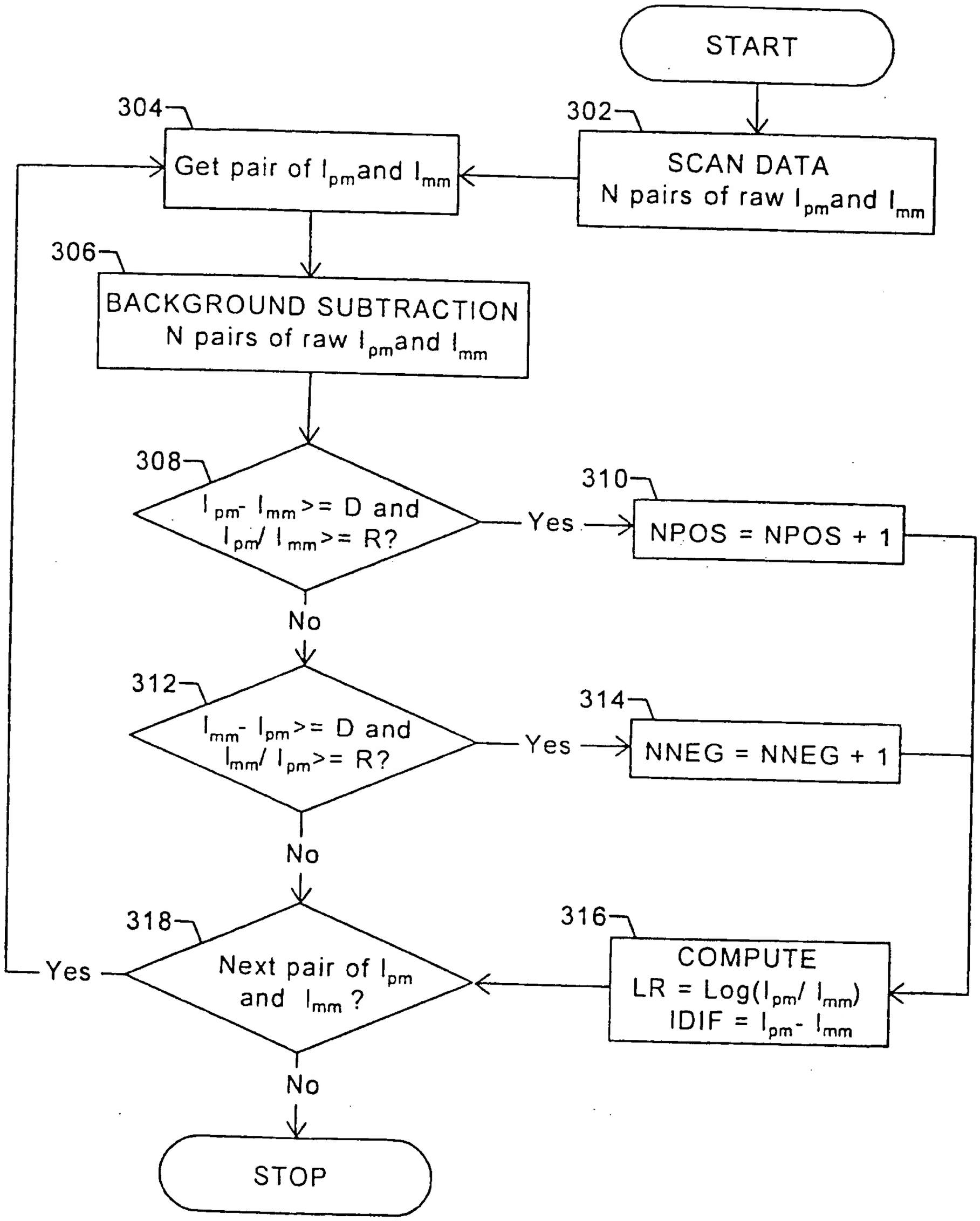
START
302
SCAN DATA
N pairs of raw I_{pm} and I_{mm}
304
Get pair of I_{pm} and I_{mm}
306
BACKGROUND SUBTRACTION
N pairs of raw I_{pm} and I_{mm}
308
$I_{pm} - I_{mm} >= D$ and $I_{pm} / I_{mm} >= R$?
Yes
310
NPOS = NPOS + 1
No
312
$I_{mm} - I_{pm} >= D$ and $I_{mm} / I_{pm} >= R$?
Yes
314
NNEG = NNEG + 1
No
316
COMPUTE
$LR = Log(I_{pm} / I_{mm})$
$IDIF = I_{pm} - I_{mm}$
318
Next pair of I_{pm} and I_{mm} ?
Yes
No
STOP

FIG. 3

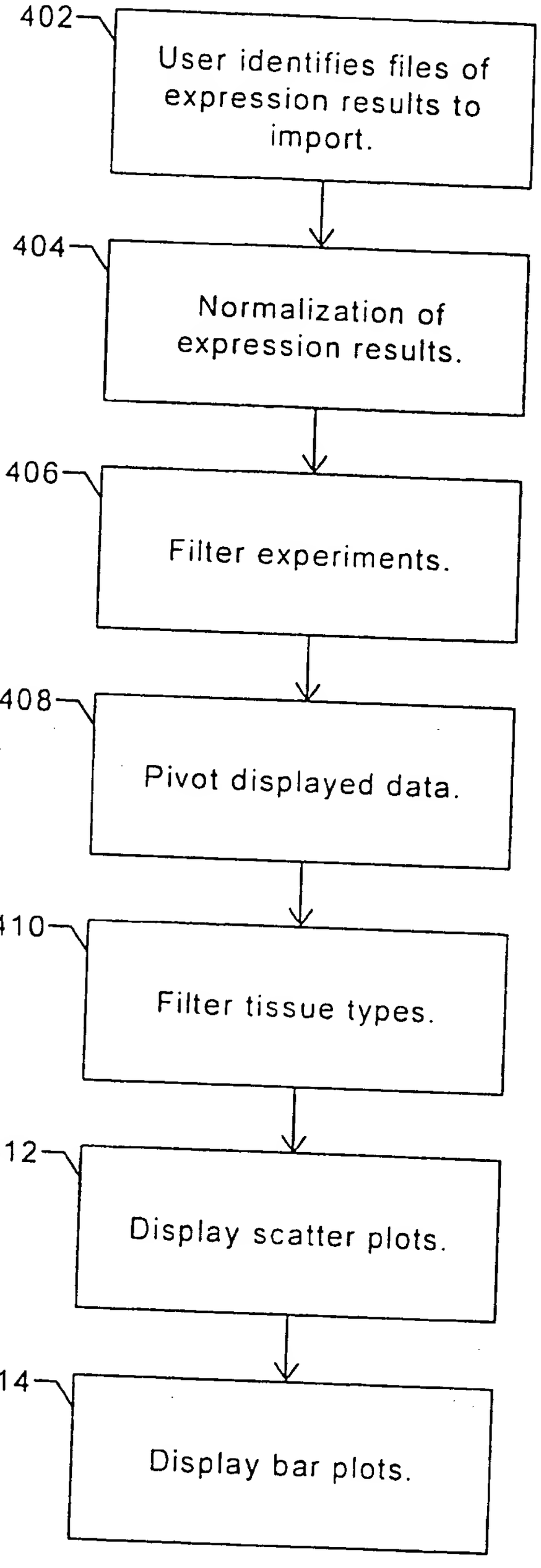
402
User identifies files of expression results to import.
404
Normalization of expression results.
406
Filter experiments.
408
Pivot displayed data.
410
Filter tissue types.
412
Display scatter plots.
414
Display bar plots.

FIG. 4

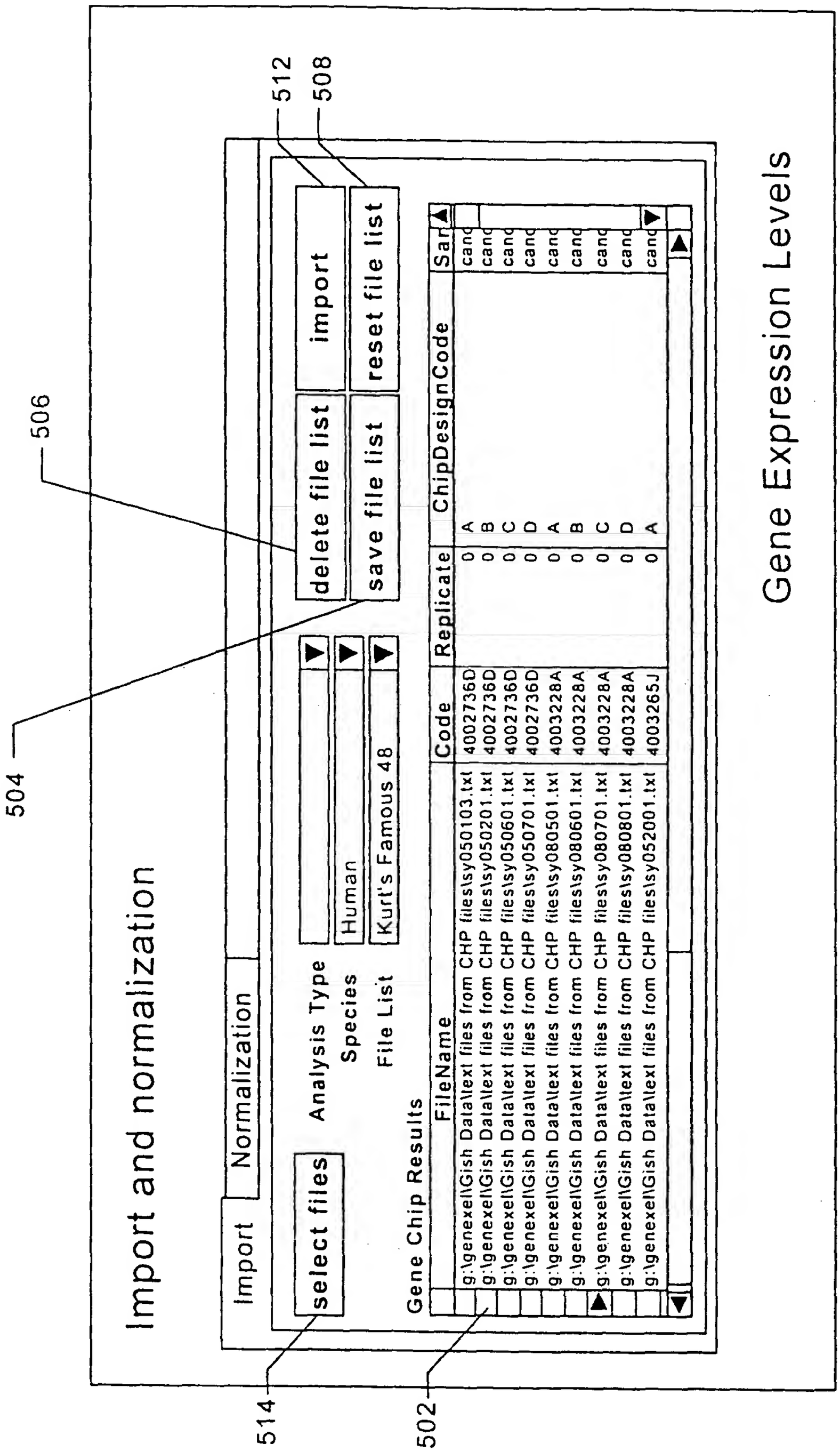
Import and normalization
Gene Expression Levels
Import
Normalization
Analysis Type
Species
File List
Gene Chip Results
select files
delete file list
save file list
import
reset file list
Human
Kurt's Famous 48
FileName
Code
Replicate
ChipDesignC
g:\genexell\Gish Data\lext files from CHP files\sy050103.txt 40027360 0 A
g:\genexell\Gish Data\lext files from CHP files\sy050201.txt 40027360 0 B
g:\genexell\Gish Data\lext files from CHP files\sy050601.txt 40027360 0 C
g:\genexell\Gish Data\lext files from CHP files\sy050701.txt 40027360 0 D
g:\genexell\Gish Data\lext files from CHP files\sy080501.txt 40032280 0 A
g:\genexell\Gish Data\lext files from CHP files\sy080601.txt 40032280 0 B
g:\genexell\Gish Data\lext files from CHP files\sy080701.txt 40032280 0 C
g:\genexell\Gish Data\lext files from CHP files\sy080801.txt 40032280 0 D
g:\genexell\Gish Data\lext files from CHP files\sy052001.txt 40032650 0 A
502
504
506
508
512
514

FIG. 5A

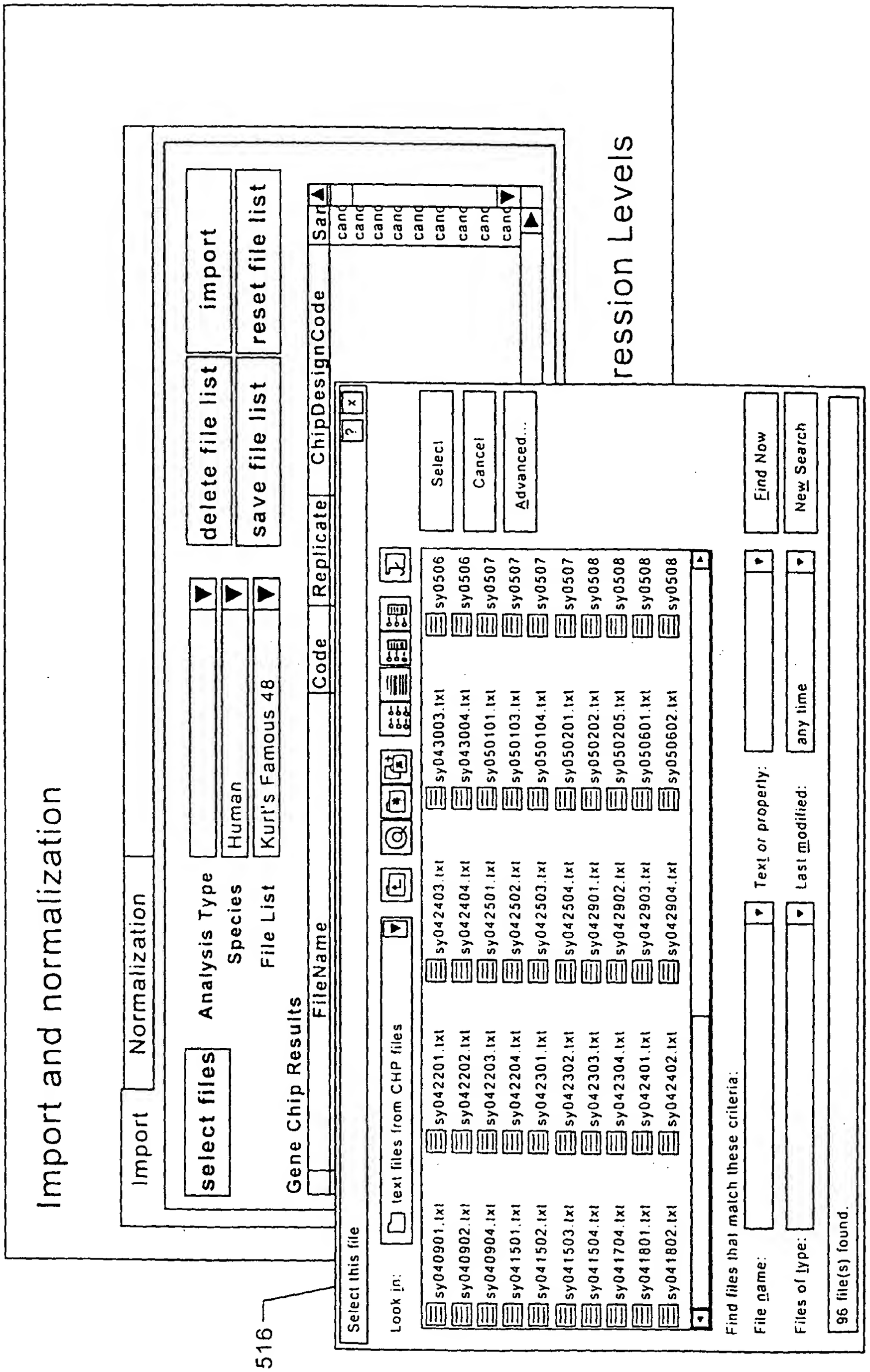
Import and normalization
Import
Normal
select files
Analysis Type
Species
Human
File List
Kurt's Famous 48
Gene Chip Results
import
reset file list
delete file list
save file list
Code
Replicate
ChipDesi
Select
Cancel
Advanced...
Find Now
New Search
Text or property:
Last modified:
any time
Select this file
Look in:
text files from CHP files
sy040901.lxl
sy040902.lxl
sy040904.lxl
sy041501.lxl
sy041502.lxl
sy041503.lxl
sy041504.lxl
sy041704.lxl
sy041801.lxl
sy041802.lxl
sy042201.lxl
sy042202.lxl
sy042203.lxl
sy042204.lxl
sy042301.lxl
sy042302.lxl
sy042303.lxl
sy042304.lxl
sy042401.lxl
sy042402.lxl
Find files that match these criteria:
File name:
Files of type:
96 file(s) found.
516

FIG. 5B

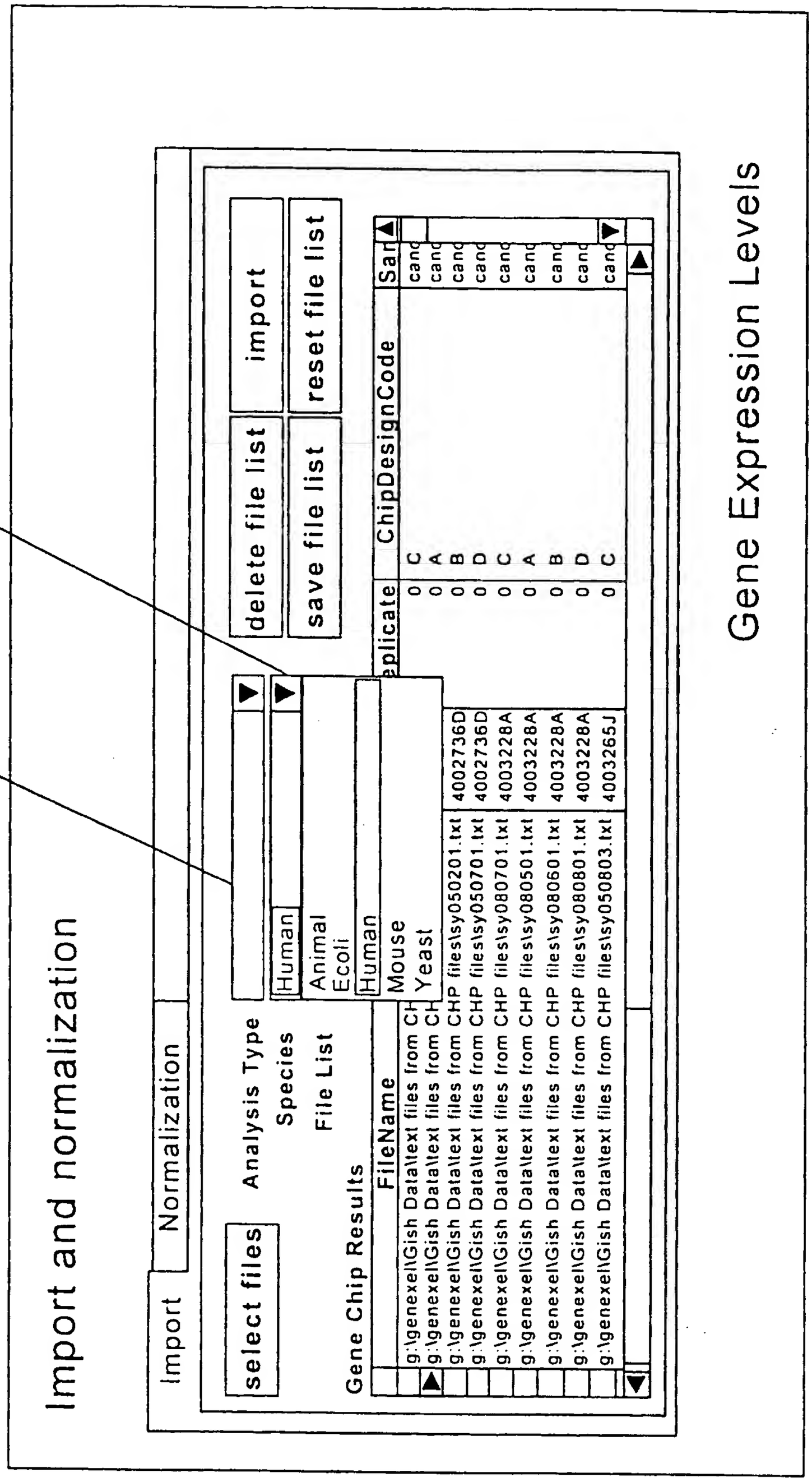
Import and normalization
Import
Normalization
select files
Analysis Type
Species
File List
Gene Chip Results
FileName
g:\genexell\Gish Data\text files from CH
Gene Expression Levels
519
518
import
reset file list
delete file list
save file list
Sar
DesignCode
ChipDe
Human
Animal
Ecoli
Human
Mouse
Yeast
40027360
40027360
40032280
40032280
40032280
40032280
40032650

FIG. 5C

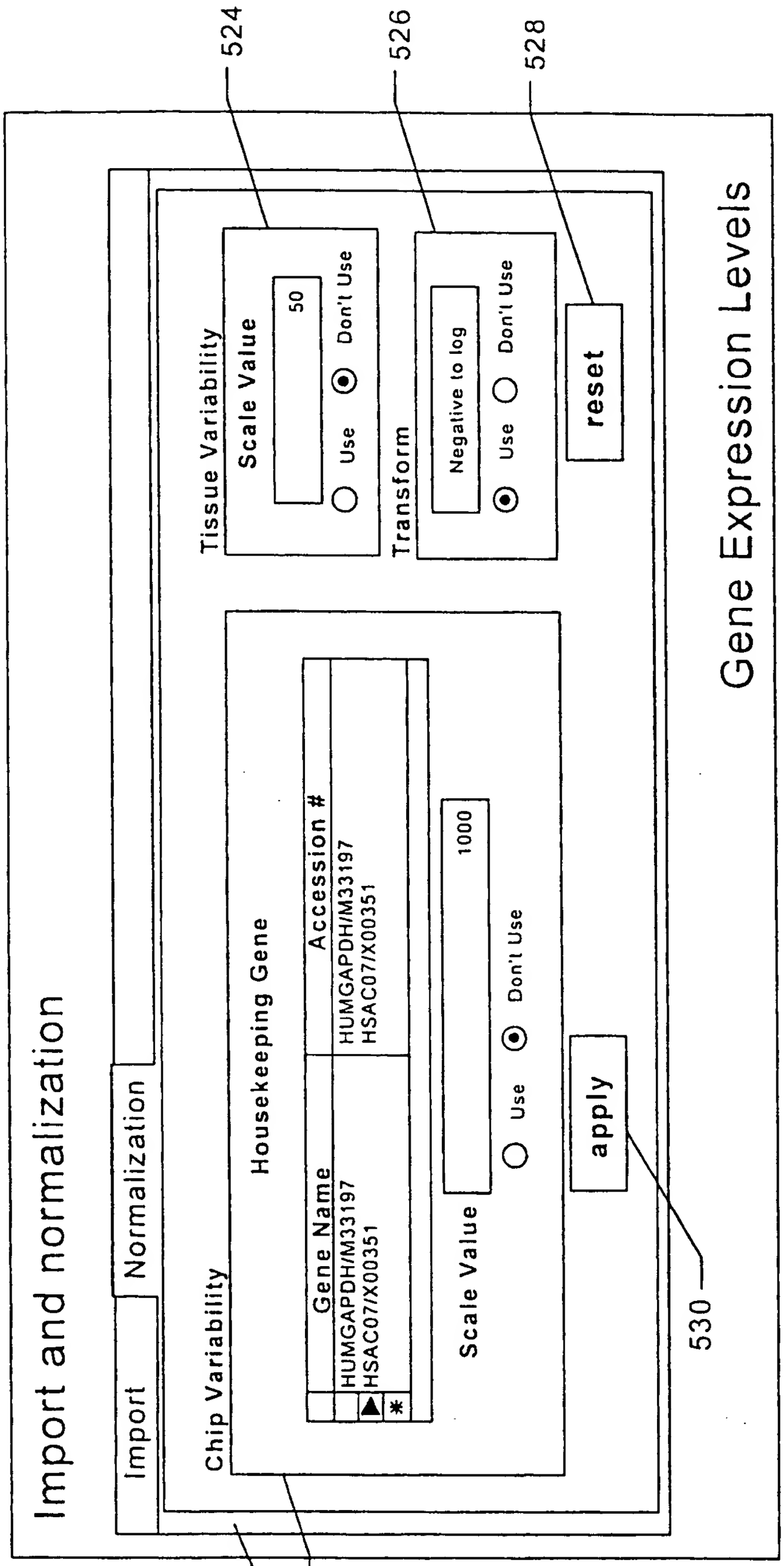
524
526
528
Import and normalization
Import
Normalization
Chip Variability
Scale Value
50
Use
Don't Use
Housekeeping Gene
Accession #
HUMGAPDH/M33197
HSAC07/X00351
Gene Name
HUMGAPDH/M33197
HSAC07/X00351
Scale Value
1000
Use
Don't Use
Tissue Variability
Scale Value
50
Use
Don't Use
Transform
Negative to log
Use
Don't Use
reset
apply
Gene Expression Levels
520
522
530

FIG. 5D

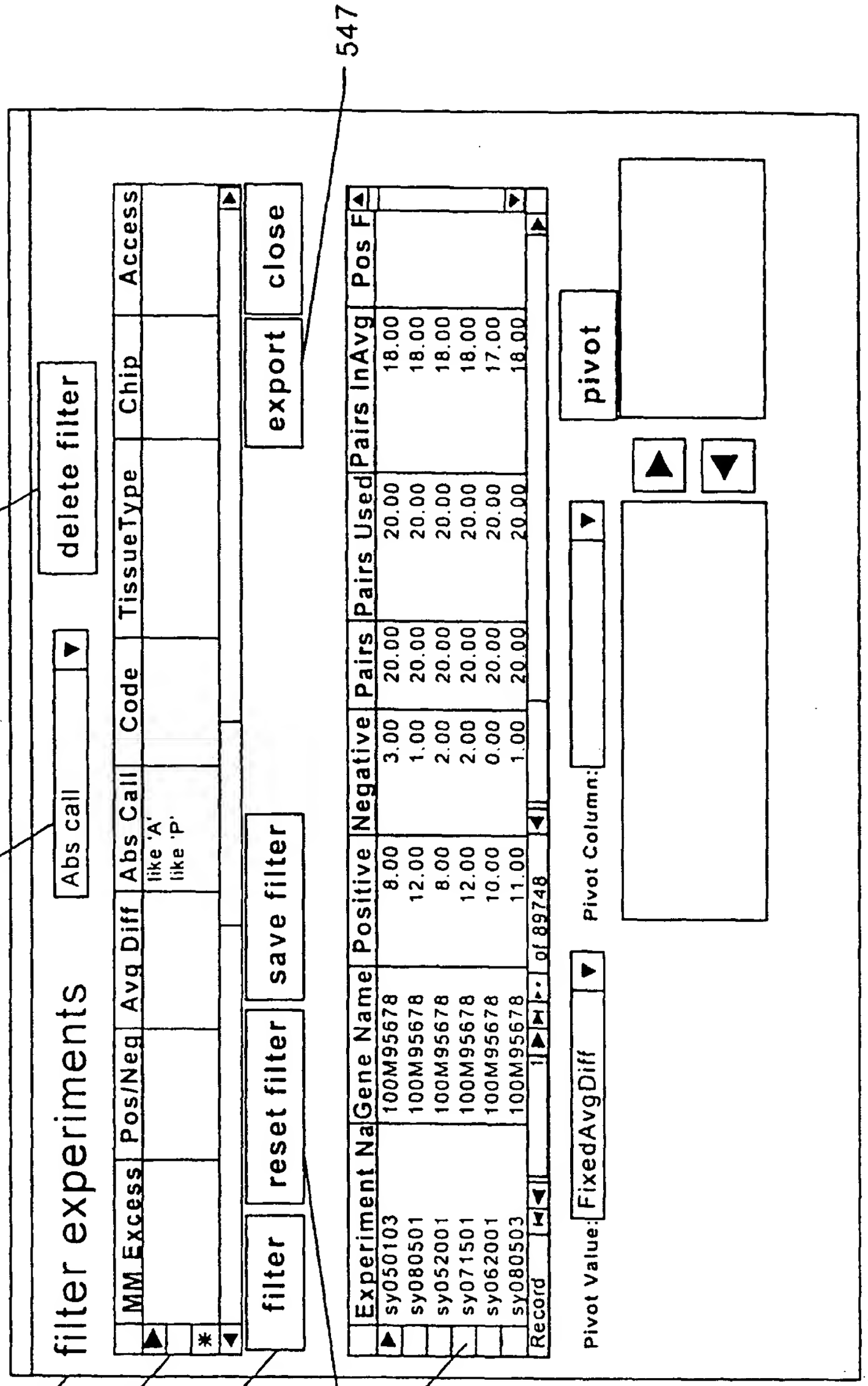
547
544
546
filter experiments
filter
delete filter
Abs call
like 'A'
like 'P'
save filter
reset filter
filter
close
export
Access
Chip
Code
TissueType
Abs Call
Avg Diff
Pos/Neg
MM Excess
Pos F
Pairs InAvg
18.00
18.00
18.00
18.00
17.00
18.00
pivot
Positive
Negative
Pairs
Pairs Used
Gene Name
Experiment Na
sy050103
sy080501
sy052001
sy071501
sy062001
sy080503
100M95678
of 89748
Record
Pivot Column:
Pivot Value:
FixedAvgDiff
532
536
538
542
534

FIG. 5E

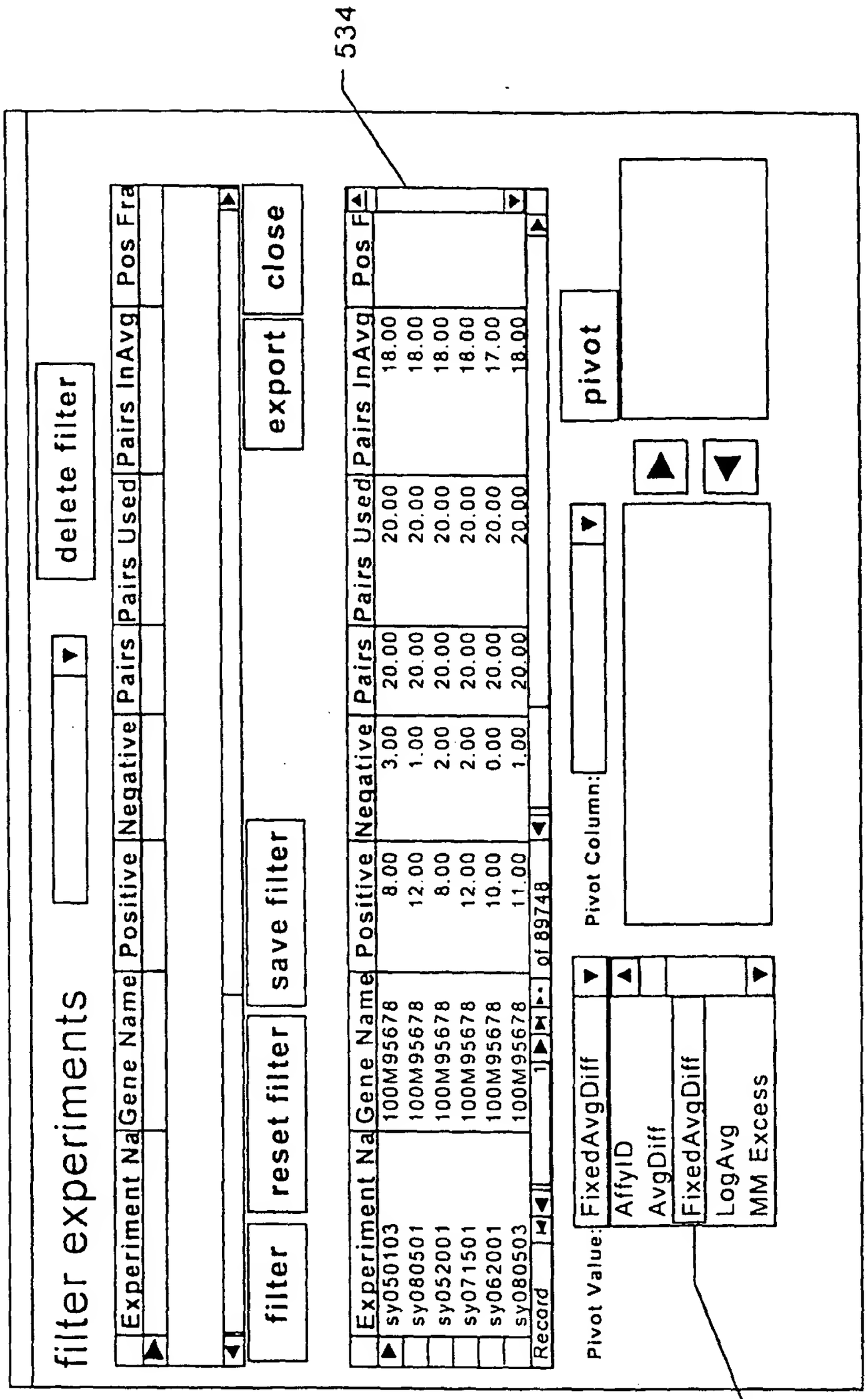
534
filter experiments
Experiment Na
Gene Nam
Positive
Negative
Pairs
Pairs Used
Pair
filter
delete filt
save filter
reset filter
close
export
Pos Fra
Pos F
airs InAvg
18.00
18.00
18.00
18.00
17.00
18.00
pivot
8.00
12.00
8.00
12.00
10.00
11.00
3.00
1.00
2.00
2.00
0.00
1.00
20.00
20.00
20.00
20.00
20.00
20.00
20.00
20.00
20.00
20.00
20.00
20.00
of 89748
sy050103
sy080501
sy052001
sy071501
sy062001
sy080503
100M95678
Record
Pivot Column:
Pivot Value:
FixedAvgDiff
AffyID
AvgDiff
FixedAvgDiff
LogAvg
MM Excess
548

FIG. 5F

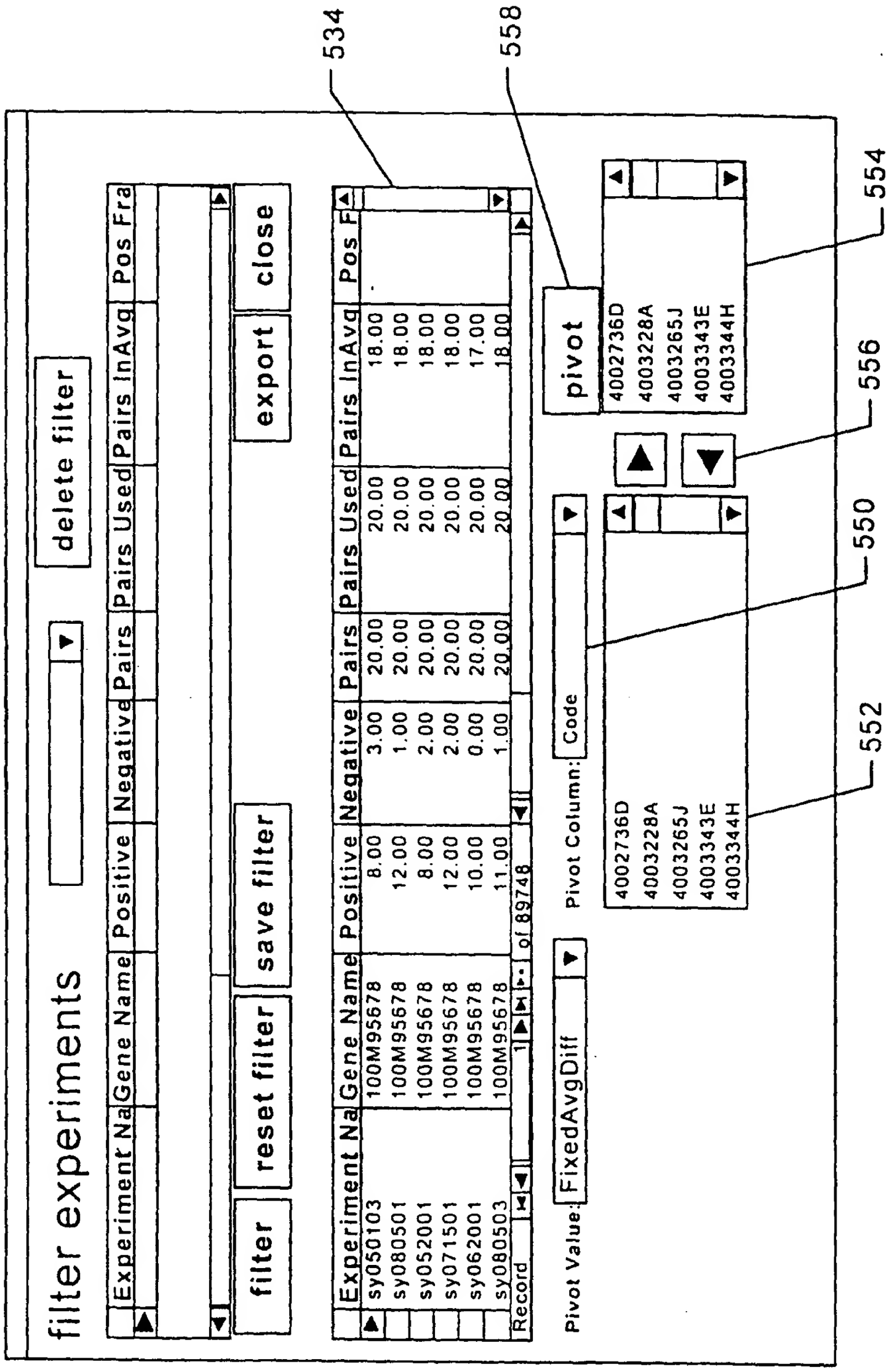
534
558
554
556
550
552
filter experiments
delete filter
save filter
export
filter
reset filter
pivot
Pivot Column:
Code
Pivot Value:
FixedAvgDiff
Experiment Na
Gene Nam
Positive
Negative
Pairs
Pairs Used
Pairs InAvg
100M95678
sy050103
sy080501
sy052001
sy071501
sy062001
sy080503
4002736D
4003228A
4003265J
4003343E
4003344H
Record
of 89748

FIG. 5G

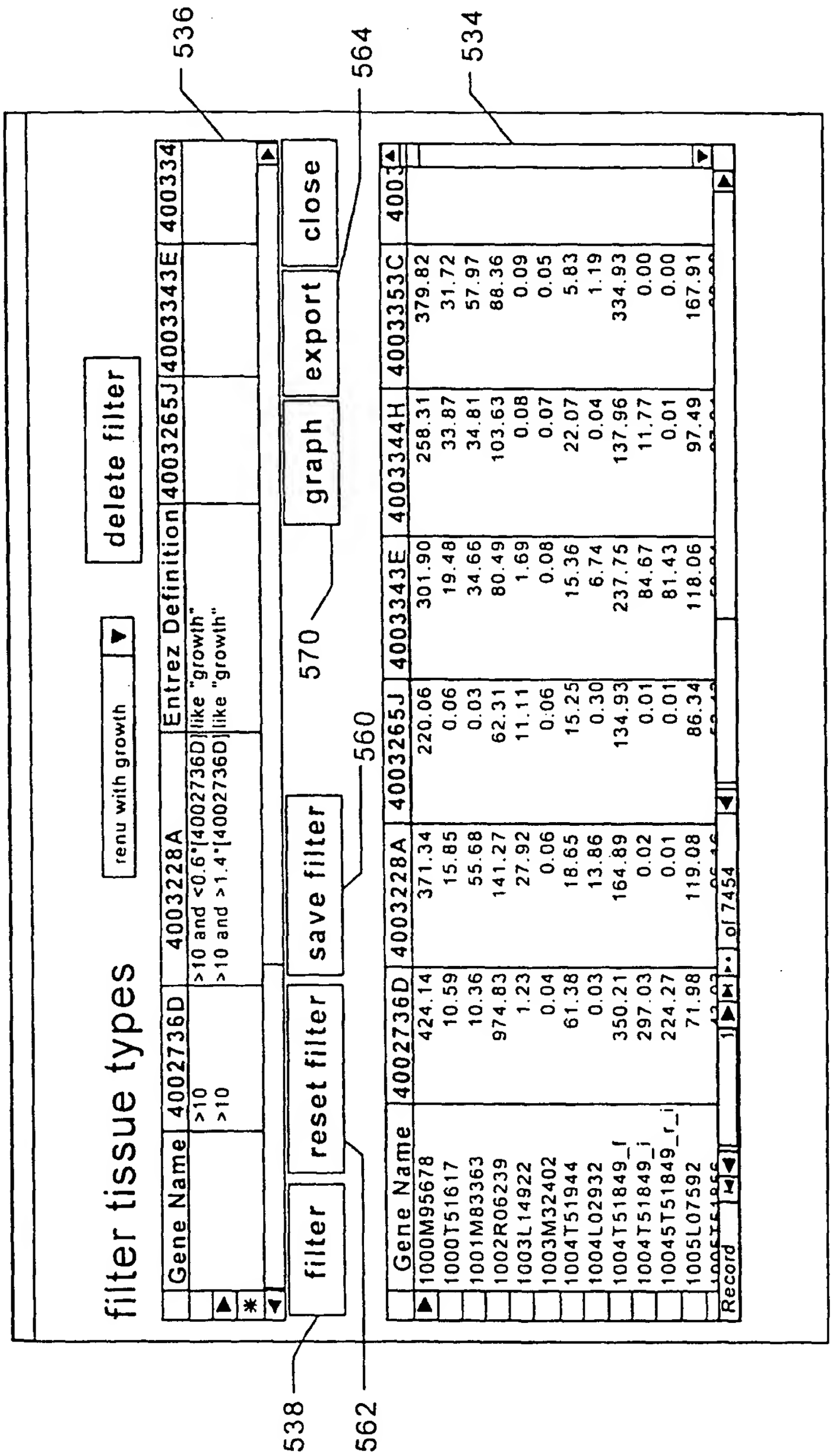
536
564
534
filter tissue types
renu with growth
delete filter
export
close
graph
save filter
filter
reset filter
570
560
538
562
Entrez Definition
like "growth"
like "growth"
Gene Name
4002736D
4003228A
4003265J
4003343E
4003344H
4003353C
400334
400334E
400334J
400334H
>10 and <0.6*[4002736D]
>10 and >1.4*[4002736D]
>10
>10
1000M95678
1000T51617
1001M83363
1002R06239
1003L14922
1003M32402
1004T51944
1004L02932
1004T51849_f
1004T51849_i
1004T51849_r_i
1005L07592
424.14
10.59
10.36
974.83
1.23
0.04
61.38
0.03
350.21
297.03
224.27
71.98
371.34
15.85
55.68
141.27
27.92
0.06
18.65
13.86
164.89
0.02
0.01
119.08
220.06
0.06
0.03
62.31
11.11
0.06
15.25
0.30
134.93
0.01
0.01
86.34
301.90
19.48
34.66
80.49
1.69
0.08
15.36
6.74
237.75
84.67
81.43
118.06
258.31
33.87
34.81
103.63
0.08
0.07
22.07
0.04
137.96
11.77
0.01
97.49
379.82
31.72
57.97
88.36
0.09
0.05
5.83
1.19
334.93
0.00
0.00
167.91
of 7454
Record

FIG. 5H

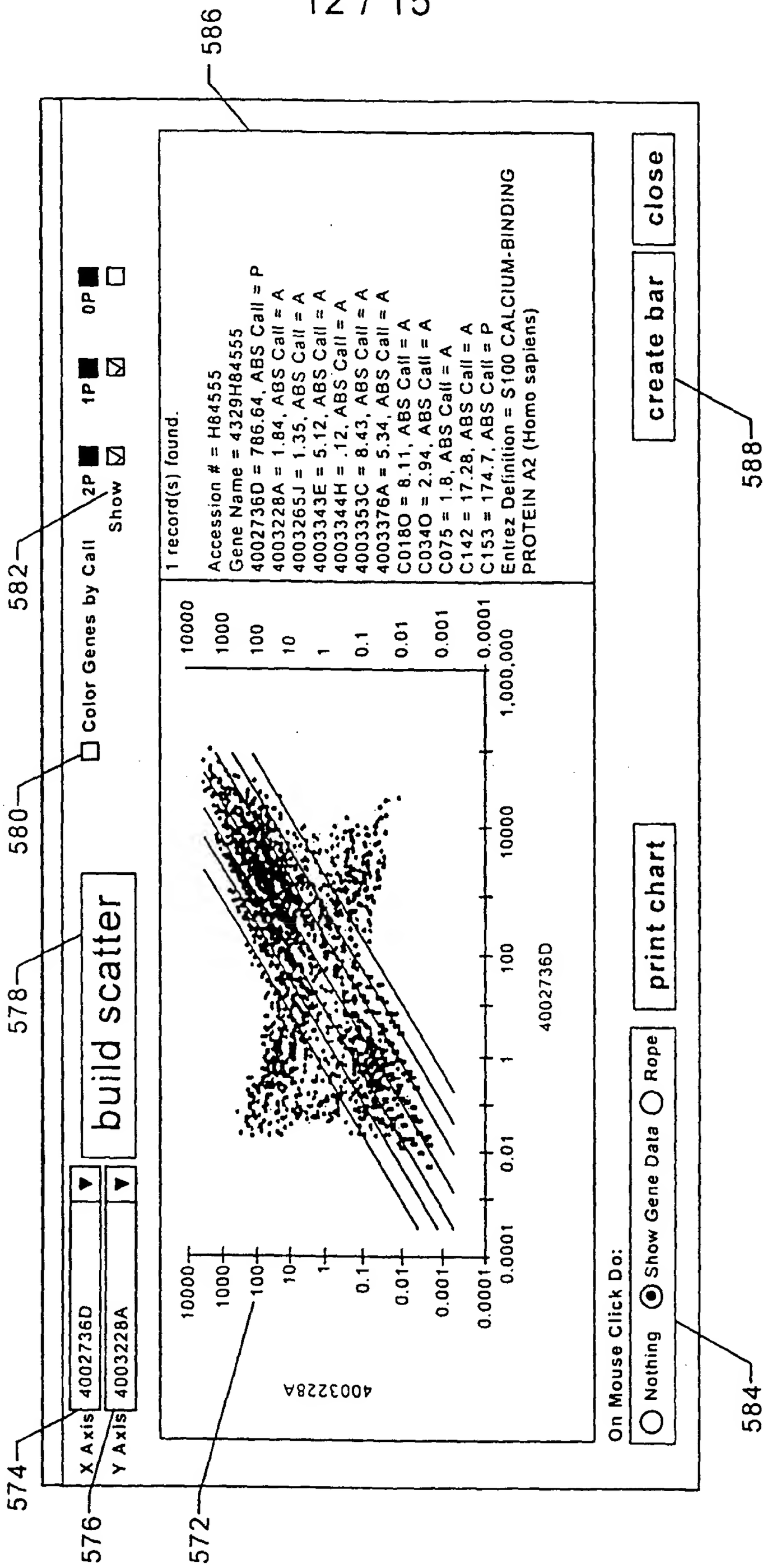
586
582
580
578
574
576
572
584
588
2P
1P
0P
Color Genes by Call
Show
build scatter
create bar
close
print chart
1 record(s) found.
Accession # = H84555
Gene Name = 4329H84555
4002736D = 786.64, ABS Call = P
4003228A = 1.84, ABS Call = A
4003265J = 1.35, ABS Call = A
4003343E = 5.12, ABS Call = A
4003344H = .12, ABS Call = A
4003353C = 8.43, ABS Call = A
4003376A = 5.34, ABS Call = A
C018O = 8.11, ABS Call = A
C034O = 2.94, ABS Call = A
C075 = 1.8, ABS Call = A
C142 = 17.28, ABS Call = A
C153 = 174.7, ABS Call = P
Entrez Definition = S100 CALCIUM-BINDING PROTEIN A2 (Homo sapiens)
10000
1000
100
10
1
0.1
0.01
0.001
0.0001
1,000,000
4002736D
4003228A
X Axis
Y Axis
4002736D
4003228A
On Mouse Click Do:
Nothing
Show Gene Data
Rope

FIG. 5I

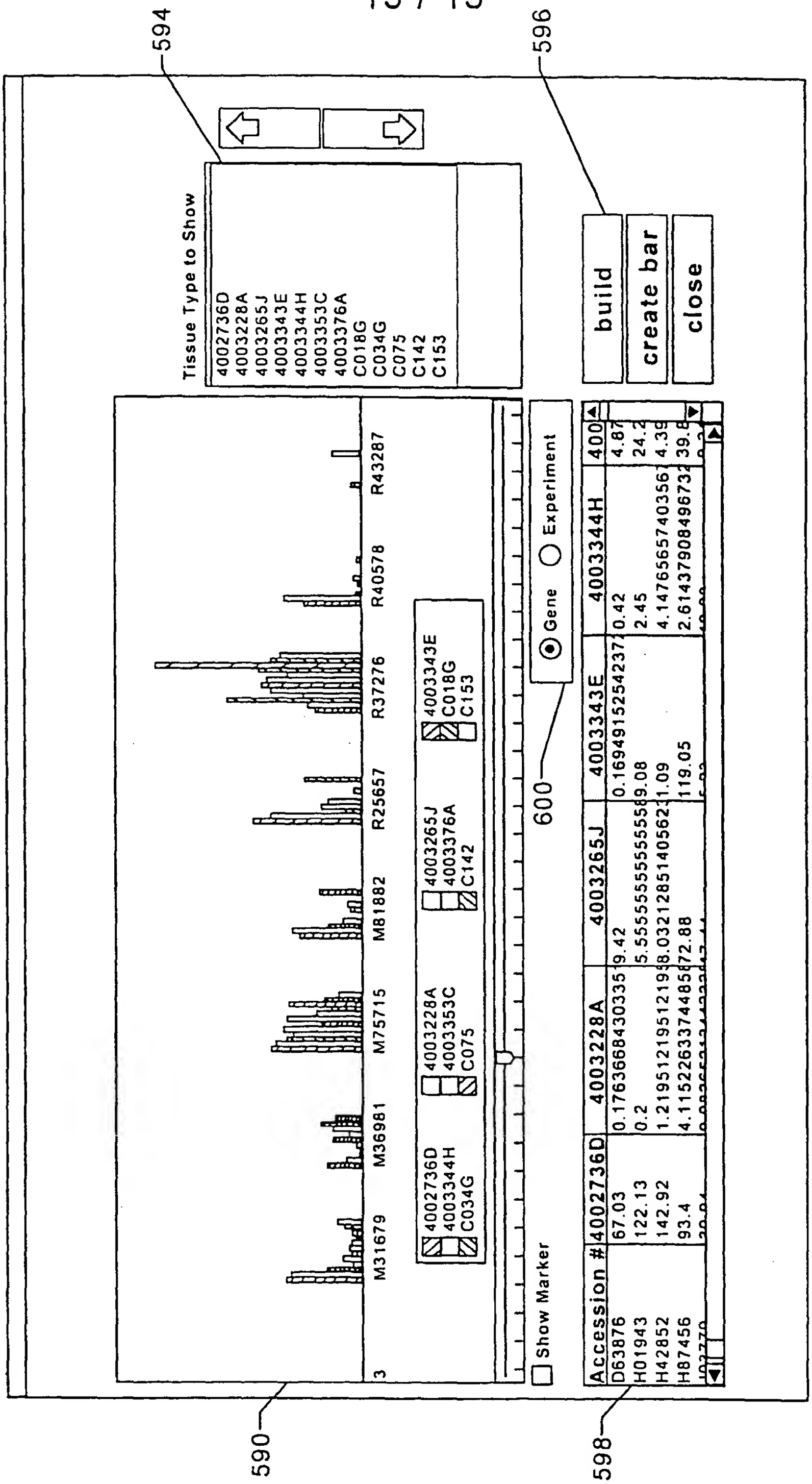
594
596
Tissue Type to Show
4002736D
4003228A
4003265J
4003343E
4003344H
4003353C
4003376A
C018G
C034G
C075
C142
C153
build
create bar
close
M81882
R25657
R37276
R40578
R43287
M31679
M36981
M75715
4003265J
4003376A
C142
4003343E
C018G
C153
4003228A
4003353C
C075
4002736D
4003344H
C034G
600
Gene
Experiment
Show Marker
Accession #
4002736D
4003228A
4003344H
4003343E
4003265J
590
598

FIG. 5J

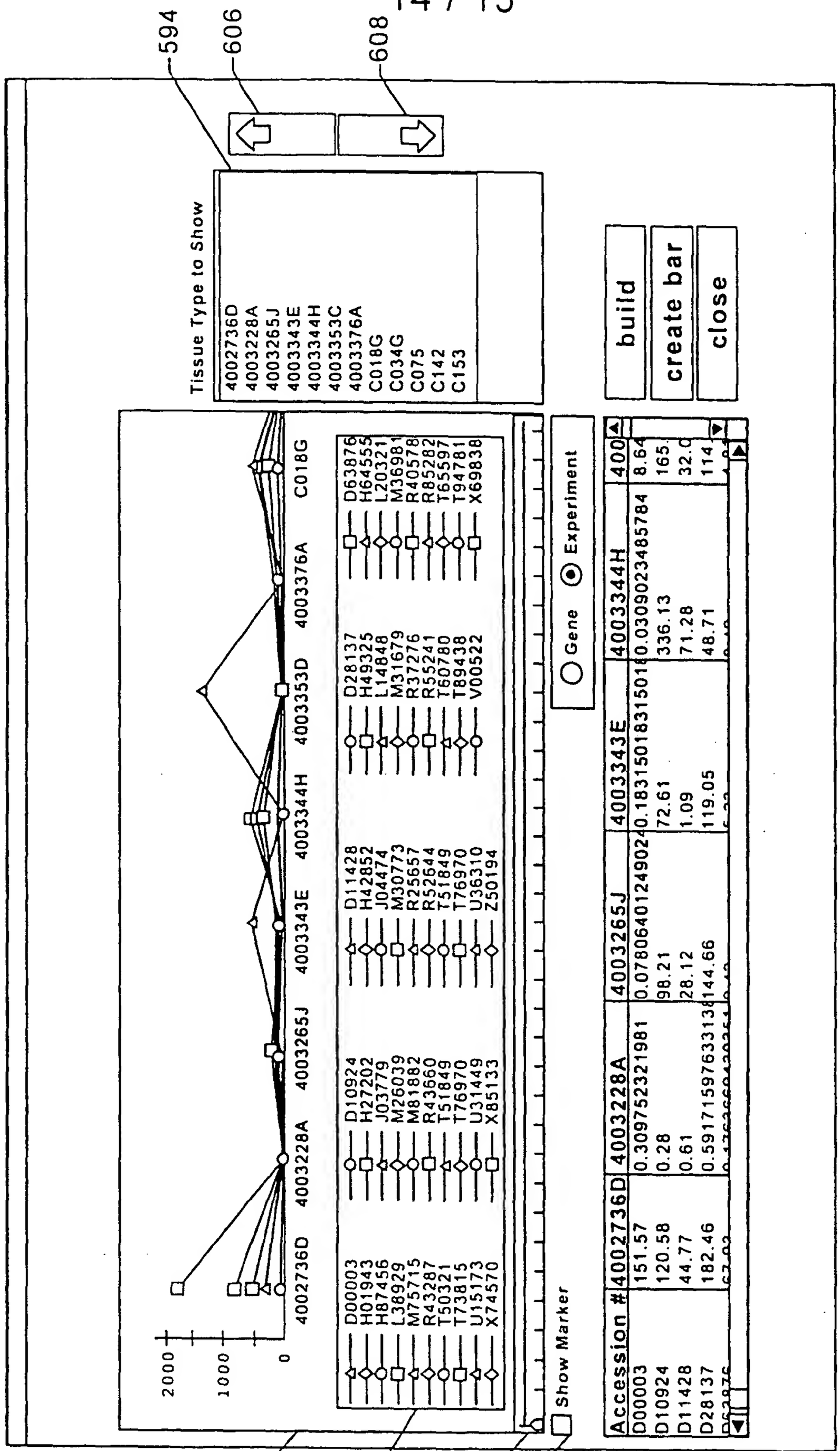
594
606
608
Tissue Type to Show
40027360
40032280
40032650
40033430
40033440
40033530
40033760
C018G
C034G
C075
C142
C153
build
create bar
close
4002736D
4003228A
4003265J
4003343E
4003344H
4003353D
4003376A
C018G
D11428
H42852
J04474
M30773
R25657
R52644
T51849
T76970
U36310
Z50194
D28137
H49325
L14848
M31679
R37276
R55241
T60780
T89438
V00522
D63876
H64555
L20321
M36981
R40578
R85282
T65597
T94781
X69838
Gene
Experiment
2000
1000
0
D00003
H01943
H87456
L38929
M75715
R43287
T50321
T73815
U15173
X74570
D10924
H27202
J03779
M26039
M81882
R43660
T51849
T76970
U31449
X85133
Show Marker
Accession #
4002736D
4003228A
4003265J
4003343E
4003344H
602
604
612
614

FIG. 5K

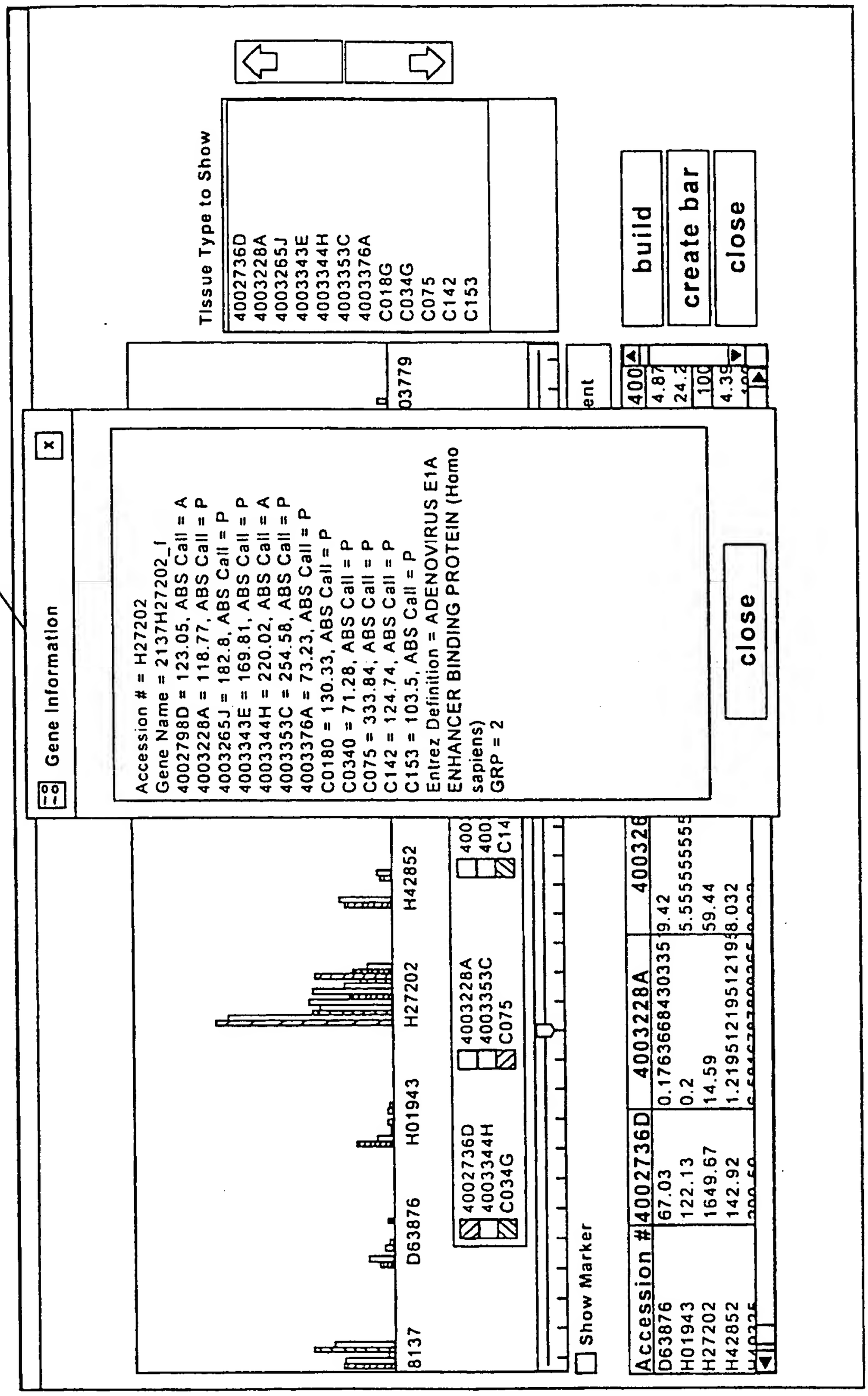
610
Tissue Type to Show
4002736D
4003228A
4003265J
4003343E
4003344H
4003353C
4003376A
C018G
C034G
C075
C142
C153
build
create bar
close
Gene Information
Accession # = H27202
Gene Name = 2137H27202_f
4002798D = 123.05, ABS Call = A
4003228A = 118.77, ABS Call = P
4003265J = 182.8, ABS Call = P
4003343E = 169.81, ABS Call = P
4003344H = 220.02, ABS Call = A
4003353C = 254.58, ABS Call = P
4003376A = 73.23, ABS Call = P
C0180 = 130.33, ABS Call = P
C0340 = 71.28, ABS Call = P
C075 = 333.84, ABS Call = P
C142 = 124.74, ABS Call = P
C153 = 103.5, ABS Call = P
Entrez Definition = ADENOVIRUS E1A ENHANCER BINDING PROTEIN (Homo sapiens)
GRP = 2
close
H42852
H27202
D63876
H01943
8137
4003228A
4003353C
C075
4002736D
4003344H
C034G
Show Marker
Accession #
4002736D
4003228A
D63876
H01943
H27202
H42852

FIG. 5L

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/15151

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12Q 1/68; G06F 17/30, 159/00

US CL :435/6; 707/1; 128/923

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 707/1; 128/923

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, STN. Search terms: transcript, transcripts, cdna#, mrna#, est#, frequenc?, abundanc?, distribut?, computer#, algorithm#

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X - Y	OKUBO et al. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression, Nature Genetics. November 1993, Vol. 2, No. 3, pages 173-179, see entire document.	1, 2, 4, and 8-10 ——— 3, 5-7, and 11-27
Y	IntelliGenetics Suite (TM), Release 5.4, Advanced Training Manual, January 1993, published by IntelliGenetics, Inc., 700 East El Camino Real, Mountain View, California 94040, USA, pages (1-6) - (1-19) and (2-9) - (2-14), see entire document.	3, 5-7, and 11-27
X - Y	WO 96/23078 A1 (INCYTE PHARMACEUTICALS, INC.) 01 August 1996, see entire document.	1-10 and 13-27 ——— 11 and 12

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search	Date of mailing of the international search report
16 OCTOBER 1998	29 OCT 1998
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer JAMES MARTINELL for Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)★